The deconstruction of a text: the permanence of the generalized Zipf law- the inter-textual relationship between entropy and effort amount

Thierry Lafouge, Abdellatif Agouzal, Geneviève Boidin-Lallich

The deconstruction of a text: the permanence of the generalized Zipf law—the inter-textual relationship between entropy and effort amount

Thierry Lafouge ; Abdellatif Agouzal ; Genevieve Lallich Boidin

1 Context and Introduction

To our knowledge, research on Zipf's law was mentioned for the first time (Petruszewycz, 1973) by *J.-B Estoup* (Estoup, 1916) and is still valid. In fact, when we searched for the "Zipf's law" character chain in the titles of articles and when we conducted a WOS query in February 2015, 70 relevant articles showed up, published since 2008. These articles came from various disciplines: physics, mathematics, economics, biology, geography. The law's original formulation is as follows: if any object in a corpus, such as a word in a text, can be characterised by a positive integer, obtained by counting, and called 'frequency', it is always possible to assign a rank to each object. *George K. Zipf* (Zipf, 1949), an American linguist, then showed that rank and frequency were not independent and approximately verified the relationship:

$$rank \times frequency \approx constant$$

This simple relationship, the original formulation of Zipf's law for a text, does not depend on any parameter. We could say that it is a paradigm (Benguigui, Blumenfeld-Lieberthal, 2011), or, perhaps, an axiom. *Egghe* classified this type of informetric law in the category of "Universal laws, not dependent on any distribution or parameter" (Egghe, 2013). The general form of this law is slightly more complex and involves a parameter, namely an exponent β positive close to 1 in its formulation (see equation (1)).

$$f(r) \approx \frac{k}{r^{\beta}} \qquad \beta > 0 \qquad r = 1,..S \tag{1}$$

where $f(r)$ denotes the frequency of the word of rank r, S the number of distinct words in the corpus (also known as the size of the text's lexicon or number of sources in informetrics) and k is a positive constant. In this article, equation (1) serves as a model to adjust the word distribution frequency, where β is a number between 1 and 2. We call equation (1) the generalized Zipf law. We won't confuse this law with Zipf's shifted law, as proposed by Mandelbrot[1], which is also sometimes known as the generalized Zipf law.

The properties of these "rank-frequency" distributions have been observed and studied in many fields. In Informetrics, the place of Zipf's law has been studied in distributions, which are often called Lotkaian in reference to the work of *Lotka* in 1926 (Lotka, 1926). These distributions model the regularities observed in the process of production or in the usage of information. A complete study of these laws can be found in (Egghe, 2005), as well as a study of Zipf's law's place among other informetric laws (see chapter 2, paragraph 4: "The place of the law of Zipf in Lotkaian informetrics").

[1] $f(r) \approx \frac{k}{(r+B)^{\beta}} \quad \beta \approx 1 \quad B > 0$

In this article, we only consider texts in which the objects are words or chains of characters[2] and the frequency is their number of occurrences in the text. We place ourselves in a particular position, if we compare it to other examples. If we reason with the formalism (Egghe, 1990) of the IPP (Information Production Process) where sources (words, researchers, cities, species, etc.) produce items (word frequency, number of articles published, number of inhabitants, number of specimens, etc.), the self-similarity of the pair (word-frequency of the word in the text) [3] can be troubling. It justifies the *Mandelbrot* fractal approach.

The first question relates to the explanation of this law and, of course, its relationship with language and writing. How can we explain it when it seems to be a universal constant (the parameter β is always close to 1 and varies little from one text to another)? The nature of this regularity is called into question with the controversial example of monkeys who randomly typed letters and pushed on spacebars, thus producing a random sequence of words (Mitzenmacher, 2003) and still verifying Zipf's law. However, if we randomly generate a text, a recent study (Ferrer i Cancho and Elveag, 2010) seems to demonstrate that there is a difference between a random text and a text written in a specific language. The authors conclude that "Zipf's law might in fact be a fundamental law in natural language". We do not provide a definite conclusion to this debate. In fact, the notion of random text is not as simple as it seems. We must first remember the various theories explaining this phenomenon, and the many different types of models (Piantadosi, 2014) that have also been offered to explain it. We shall cite the three most common approaches in informetrics.

1) *Herbert A. Simon* (Simon, 1955) described a stochastic process which produces these stationary laws. He used the example of the writing of a book to illustrate his demonstration. This describes a process to generate a text. Many demonstrations are possible to explain different phenomena with this type of process. The "cumulative advantages" law (Price, 1976) falls into this category.

2) *Mandelbrot* (Mandelbrot, 1953) argued his view differently. He considered the generation of language as the transmission of a signal and he used information theory. He showed that, in seeking to minimise the cost of producing words (the number of characters, also known as the size of the alphabet, plays a key role), the results of a generalized Zipf law could be obtained (see section 4.2.2). Of course, we must agree on the choice of a cost function, which plays an indirect role in the length of the words. This cost function, which we use, leads to words that are rare, thus requiring a much greater effort to be produced.

3) *Mandelbrot* later presented a second argument (Mandelbrot, 1977) of a geometric kind, which we recall in section 4.2.4 to explain this law.

We can conclude this broad state-of-the-art by saying that this law is universal and applies to all texts. Texts can be selected according to various criteria (size, language, date, subject, etc.). Such a distribution can be seen in languages that have disappeared today and that we do not yet know how to translate (Reginald, 2007).
A text is never finished, it is not limited and the beginning or end are irrelevant to Zipf's law. Zipf's law seems to apply as soon as the ratio between the number of sources (size of lexicon) and the total number of words are balanced[4].

---

2 Chains, unlike words, are devoid of meaning ; semantics are absent.

3 In fact, in this case, a word in a text is characterised by its frequency alone: variables and probabilites seem to merge.

4 We find in [Simon, 1955] paragraph 4 "The empirical distributions Word frequency" many ideas on the increased size of the lexicon: I quote p. 434 : "*An author writes not only by processes of association - i.e. sampling earlier segments of word sequence - but also by processes of imitations - i.e. sampling segments of word sequences from other works he has written, from works of onther authors...* "

2 Hypotheses and Objectives

In this article, we discuss results taken from models 2) and 3) as described above. The first Mandelbrot model is taken from a larger class of models and depends on communicative optimization principles (see section 4.4 Commutative accounts in (Piantadosi, 2014)). The second model consists in saying that Zipf's law is essentially a statistical artefact (see section 4.1 Random typing accounts in (Piantadosi, 2014]). In both of these models, the size of words is very important, but in different ways. In the first model, coding is optimized when the message is sent depending on the length of the words. It pertains to the field of language. The second model pertains more to the field of writing, in which patterns of certain lengths are repeated.

We therefore ask the following question: when the text is modified, does the regularity of word frequencies stay the same, thus making the length of words invariant? In order to verify our hypothesis, our idea consisted in taking a text, which we modified and degraded; hence the term 'deconstruction' in the article's title. To be more precise, our method consisted in applying an algorithm that replaced characters one by one with a "joker' in the whole text through successive iteration. We then segmented the modified text into words. We therefore didn't change the distribution of word lengths in the text. We then checked whether a generalized Zipf law (see (1)) was still valid. We found that it was constant with a different β coefficient that increased with each step. However, a degraded text is no longer a text. The segmented sources are no longer words belonging to a text or to a specific language!
In the final step, all of the characters were deleted (the text became a string of jokers and separators). The distribution of frequencies and the distribution of word lengths merged. However, the distribution of lengths was not zipfian. What happened during this deconstruction?

Also, the fact that we can always adjust the distribution frequency of sources during the first steps raises new questions. This result is not incompatible with *Mandelbrot*'s second demonstration as long as the size of the alphabet is not too small. It is also possible to process these 'degraded texts' with Shannon's signal theory. In that case, we must calculate the degraded text's entropy and its associated amount of effort. If Mandelbrot's hypothesis concerning language transmission is true, there must be a trace of it during the deconstruction phase. Entropy and the amount of effort cannot be independent.
This allows us to see that entropy and the amount of effort are not independent and that inter-text regularities can be highlighted. This is why we subtitled our article as "the inter-textual relationship between entropy and effort amount". We do not aim to build a new explanatory model, since many of them already exist and are very relevant. Instead, we want to shed new light on inter-textual statistical regularities and build on Mandelbrot's work.

Our article is organised in four parts:

- The method of depleting a text (see section 3) is first explained.
- We then recall the mathematical tools and concepts used: entropy, amount of effort. These results show that, for degraded texts, a relationship exists between entropy and the amount of effort at each step (see section 4). This section develops Mandelbrot's theoretical results and applies them to a degraded text.
- Our method is applied on a text of 114,730 words segmented into 12,740 sources (see section 5).
- Thanks to the latter experiment, we are able to show that there is a linear relationship between the amount of effort and entropy (see section 6). The theoretical results (section 4) that were previously found are insufficient to explain the experimental results in a

satisfactory way. A possible explanation is offered, using recent works (Lafouge and Agouzal 2015) that combine power laws and effort functions.

Finally, we end with a conclusion.

3 The Depleting Method

3.1 Lexical Definitions

A text is one of the tangible forms that stems from intellectual activity, namely writing. A text is a set of signs, ordered and grouped into words. It is therefore composed of a sequence of ordered words. Before defining what we call a "word", we must clarify a few definitions. We define signs within a text, namely all characters and separators, as graphemes belonging to the whole and consisting of two disjointed subsets.

- By characters, we mean elements of the language's alphabet: letters (uppercase and lowercase) and numbers.
- We distinguish three types of separators:
  - Physical separators: blanks and line breaks
  - Punctuation separators: full stops, commas, exclamation marks, and quotation marks
  - Special signs: open brackets, closed brackets

The 9 separators that we retained for the experiment are:
Blanks, line breaks, open brackets, closed brackets, exclamation marks, question marks, full stops, commas and quotation marks.
This list of separators is arbitrary and other choices could have been made.

We define a "word" as being a sequence of characters placed between two separators. The definition of the separator in subscript allows the computer to identify words automatically. Physical separators generally suffice for experiments using Zipf's law. This choice has very little effect on the results. However, we can regret the absence of clarification by authors when verifying the validity of *Zipf's* law. In the following experiment, we pre-processed uppercases. All uppercase letters were replaced by lowercases, except for proper noun initials. This can be done automatically if proper nouns are marked.

3.2 An Algorithm for the Deconstruction of a Text

Characters are substituted one by one in the text, and replaced by a sign that does not exist in the language, known as the joker and written as @. Our algorithm is illustrated with the sentence: " *the monkey types on the keyboard with strength.*

1. Initiation : *The monkey types on the keyboard with strength.*

2. Elimination of "e" : *Th@ monk@y typ@s on th@ k@yboard with str@ngth*

3. Elimination of "s, a, t, n, r " :*@h@ mo@k@y @yp@@ o@ @h@ k@ybo@@d wi@h @@@@@@@@h*

4. Elimination of "m, h, d, k ": *@@@ @o@@@@y @yp@@ o@ @@@ @@ybo@@@ wi@@ @@@@@@@@@*

5. Elimination of "w, i, b, o, y" : :*@@@ @@@@@@@ @@@@@ @@ @@@ @@@@@@@@@ @@@@ @@@@@@@@@*

The total number of signs is invariant, the size of the alphabet decreases, going from 16 (this is the number of iterations needed to completely deconstruct the text) to 1. The total number of words is invariant and remains equal to 8. The number of sources, equal to 7, decreases and is then equal to 5 (we will prove later on that, when deconstructing a text, the number of sources either decreases or remains constant at each step). Replacing deleted characters with a joker maintains the same distribution of length for each word throughout the process. The text structure is preserved since the proportion of separators does not change over the course of the algorithm. The joker is a simple artefact and is not a critical part of the process. In the last iteration, if we segment the text, we obtain the distribution of word lengths.

To implement this algorithm, we remove the characters by decreasing frequency. Other character substitution methods are tested in this paper, with comparable results.

3.3 Segmenting and Formatting Distributions

At each step, noted *i,* we segmented the depleted text – later written as T_i – into words. We identified the number of sources, noted S_i, and we ranked the frequency of sources – written as $f_i(r)$ – as well as their frequency in the text. Finally, to verify the generalized Zipf law, we constructed a frequency-rank distribution:

$$\big(r, f_i(r)\big) \quad r = 1..S_i \tag{2}$$

where sources are ranked by decreasing frequency.
We note that, for each iteration *i*, when a character is substituted, then the number of sources S_i decreases[5]. Let us prove this result. If we call V_i the number of distinct characters in a text at step *i* and *L* the maximum length of a chain of characters in the text, then the text's number of possible sources $Smax_i$ in step *i* is equal to the finite sum:

$$Smax_i = \sum_{j=1}^{L} (V_i)^j$$

We can easily show that $Smax_i$ decreases at each step *i* because V_i decreases by 1 at each iteration *i*. $Smax_i$ is equal to *L* when the last character has been substituted. This allows us to say that the number of sources S_i therefore also decreases at each step *i* (it may remain constant when we substitute a rare character). So S_i is a decreasing sequence.

3.4 The Adjustment of Distributions

At each deconstruction step (the text is not modified in the first iteration, since only one letter has changed), we carry out an adjustment of the distribution (2). We postulate[6] that distribution (2) is a generalized Zipf law at each step. This allows us to write:

$$f_i(r) \approx \frac{k_i}{r^{\beta_i}} \quad k_i > 0 \;\; \beta_i \geq 1 \;\; r = 1,..S_i \tag{3}$$

5 This property is true since we delete the same character throughout the text.
6 This hypothesis is verified at all steps of the deconstruction.

We limit ourselves to $\beta_i \geq 1$, which is almost always verified in practice. We know that the word frequency distribution is more complex than initially thought (see section 3, "The word frequency distribution is complex" in (Piantadosi, 2014)). For example, the log-normal law (Petruszewycz, 1972) is very similar to the inverse power-distribution. In order to calculate the adjustment parameters k_i and β_i, we carry out a simple linear regression after having transformed the coordinates on an $Ln - Ln$ scale. By calculating the coefficient R^2, we measure the quality of the adjustment. When it is greater than 0.97, we speak of an adjustment[7] with the generalized Zipf law. Many other methods exist to adjust this type of distribution. We can, for example, refer to (Clauset, Slalizi, Newman, 2009). The visualisation of the curves and classic R^2 test are sufficient for this study. In fact, we know that a very significant R^2 test can sometimes hide a distribution that is not uniform; hence the importance of viewing the graph to avoid errors.
The adjustment itself isn't the determining factor in this study, and this is why we chose not to test other models. The most interesting statistical regularities here (see section 6) don't depend on a model and depend even less on an adjustment method.

4 Results and Mathematical Treatments

4.1 Informetric Results

Let us first establish an informetric result which is a property of the generalized Zipf law (see (1)), and is proved in Annex 1. This mathematical result could have been omitted from this article as it is not fundamental. However, it can help to answer some questions that could arise during the experiment.
In what follows, we use the continuous mode. Each stage is characterised by the number of sources S_i. The number of words (written *M*) is invariant throughout the deconstruction process. If adjustment (3) is perfect, we come up with the following equality at each step of the deconstruction:

$$\int_1^{S_i} \frac{k_i}{r^{\beta_i}} dr = M$$

This amounts to writing the equation:

$$\int_1^{S_i} \frac{dr}{r^{\beta_i}} = M_i \ ; \quad M_i = \frac{M}{k_i} \tag{4}$$

The theoretical question we ask is as follows: if (S_i, M_i) is a sequence, does a unique β_i exist so that equation (4) is verified?
In Annex 1, we prove the necessary and sufficient condition needed to solve this mathematical problem and we present an additional result that explains why β_i is an increasing sequence.

4. 2 The Theoretical Relationship between Entropy and Effort Amount in the Degradation of a Text

We built on Mandelbrot's results and applied them to the deconstruction of a text. We proved the "linearity" relationship (12) between entropy and the inter-textual amount of effort.

4.2.1 Entropy and the Amount of Effort at Each Step of Deconstruction

7 For each adjustment, we display the graph with the points that are aligned to see if there is no bias.

We place ourselves within *Shannon*'s signal theory. The entropy[8] of the degraded text T_i as represented by distribution (2) is

$$H_i = -\sum_{r=1}^{S_i} Ln(\, p_i(r)).\, p_i(r) \qquad (5)$$

where $p_i(r) = \frac{f_i(r)}{S_i}$ where $p_i(r)$ is the probability of a word of rank r in the degraded text T_i.

We assume that there exists an effort function written as $C_i(r)$ that is strictly positive. This function allows us to define the total amount of effort E_i of the degraded text T_i with:

$$E_i = \sum_{r=1}^{S_i} C_i(r)\,.\, p_i(r) \qquad (6)$$

The problem is then to choose which effort function to use. We use the effort function ([*Mandelbrot*, 1953]) defined by *Mandelbrot* .

$$C_i(r) = \frac{\text{Ln}\,(r)}{\text{Ln}\,(V_i)} \qquad (7)$$

This effort function has the following characteristics:
- As the number of distinct characters is reduced, the number of characters in words increases, and, according to (7), the amount of effort needed to produce them increases.
- The higher the rank is, the more words tend to have a large amount of characters and the rarer they are. According to (7), more effort amount is needed to produce them as the rank increases.

4.2.2 Minimizing the Cost of Transmission: Mandelbrot's Optimization Model

Knowing the entropy H and the amount of effort E (see (5), (6), (7))), Mandelbrot then calculated the probability $p(r)$ of a word by minimizing the average cost of information, $C = \frac{E}{H}$ Following this calculation, he obtained the following equality (see for example the demonstration section *Power Laws via Optimization* in (Mitzenmacher 2003)):

$$p(r) = \frac{k}{r^{\beta}} \qquad \beta = \frac{1}{C.\text{Ln}\,(V)} \qquad (8)$$

We therefore obtain a proportional relationship between entropy and the amount of effort:

$$H = \beta.\text{Ln}(V)\,.\,E \qquad (8b)$$

In the case of a text's deconstruction, this hypothesis seems harder to apply to each deconstructed text: we choose not to use this result here, but we retain formula (7) to calculate the amount of effort.

4.2.3 The Link between Entropy and the Amount of Effort

[8] Here, entropy is measured by a natural logarithm; it should be divided by Ln(2) to express it in bits.

If we calculate entropy (5) assuming that the distributions are generalized Zipf law, using the equation (3),(4),(6),(7) we obtain a relationship between entropy and amount of effort:

$$H_i = \gamma_i + \delta_i . E_i \quad \gamma_i > 0, \quad \delta_i = \beta_i . \mathrm{Ln}(V_i) > 0 \qquad (9)$$

The T_i text's entropy depends on adjustment parameters, on the amount of effort and on the number of distinct characters. The proof of this result can be found in Annex 3. Later on, we modified this result using Mandelbrot's geometric study of Zipf's law.

4.2.4 The Theoretical Extension of Mandelbrot' Work: the Inter-textual Relationship between Entropy and Effort Amount

We need a strong hypothesis to demonstrate *Mandelbrot's* 1953 result (8). This hypothesis is applied to the construction of the text and consists in minimizing the cost of information when transmitting the signal.

This hypothesis is criticised by *Simon*. A controversy recounted in (Mitzenmacher, 2003) ensued between the two researchers.
Later, *Mandelbrot* supplied quite a different proof for Zipf's law. Each word is a sequence of characters framed by two separators, and is characterised by its length. The coefficient β can be interpreted as a fractal dimension. This approach is more appropriate for processes that deal with writing (Lafouge, Pouchot, 2012). Mandelbrot shows that β depends on the size of the vocabulary V; more precisely, he obtains (*Mandelbrot*, 1977]):

$$\beta = \frac{-\mathrm{Ln}\,(\rho)}{\mathrm{Ln}\,(V)} \qquad (10)$$

where ρ, a character's probability of occurrence, is a difficult parameter to measure and is not independent of *V*. We necessarily have $\rho \leq \frac{1}{V}$. This formula is difficult to verify. A detailed demonstration of (10) can be found on pages 43-44 (Egghe, 2005). This demonstration can be applied to deconstructed texts, as long as V_i is big enough. In this case, we speak of a fractal hypothesis, because $\frac{1}{\beta}$ can be interpreted as a fractal dimension.

If we choose $\rho = \frac{1}{V}$, we therefore find, according to (10) :

$$\beta = \frac{\mathrm{Ln}\,(V+1)}{\mathrm{Ln}\,(V)} \; . \qquad (10b)$$

The coefficient's formula (10b) is identical to the result of (Li, W. ,1992). In his article, *Li* shows that any random text has a word frequency distribution that verifies the generalized Zipf law offered by Mandelbrot (see footnote 1). If we assume that $V = 26$, we have, according to (10b): $\beta \approx 1{,}01$. The more the size of the alphabet increases, the closer the coefficient is to 1. Conversely, a text coded with two characters (text with 0 and 1 and with one separator) has a coefficient such that $\beta \approx 1{,}584$.
The value of the β coefficient for random texts varies between 1 and 1.6.

Equation (10) shows that, if ρ is considered constant during the deconstruction (which is plausible since the proportion of separators remains constant) then β_i increases when V_i decreases.

If ρ is considered to be constant during the deconstruction, $\beta_i.\ln(V_i)$ doesn't depend on i and is therefore constant if we have:

$$\beta_i.\mathrm{Ln}(V_i) = \delta \qquad (11)$$

Equation (9) is thus written as:

$$H_i = -\mathrm{Ln}(h_i) + \delta.E_i \qquad i = 1..I \qquad (12)$$

where I is the number of iterations for which a generalized Zipf law is always acceptable. The next experiment aims to study the relationship between entropy and effort amount during the deconstruction of a text.
Can the « quasi-linearity » relationship (12) between H_i and E_i be verified experimentally?
To summarize, proving formula (12) supposes that the following hypotheses are verified:

- (a) a generalized Zipf law at each step of the deconstruction,
- (b) an effort function equal to $C_i(r) = \frac{\mathrm{Ln}\,(r)}{\mathrm{Ln}\,(V_i)}\ r = 1,..S_i$ (function used by Mandelbrot) at each step of the deconstruction
- (c) a constant $\beta_i.Ln(V_i)$ product (the consequence of Mandelbrot's fractal hypothesis).

The following experiment will allow us to verify hypotheses (a) and to delve further into the Inter-textual relationship (12) between entropy and effort amount.

5. Application: the Deconstruction of a Text

5.1 The Text's Characteristics

The deconstructed text is *Principes de gégraphie humaine* (*Principles of Human Geography*) written in 1921 by Paul Vidal de la Blanche, a French geographer (1845-1918).[9] The lexicometric characteristics, calculated after the segmentation of the text of 212 pages, are:

- Total number of signs: 680,564
- Percentage of separators: 18.9%
- Percentage of characters: 81,1 %
- Total number of segmented words: 114,730
- Number of identified sources: 12,747
- Number of distinct characters: 82

This text is available online using the URL address displayed in the footnotes. The reader who wishes to test this algorithm can do so on any kind of text and will obtain similar results.

5.2 The Characteristics of 27 Deconstructed Texts

9 This text can be found at the following address: https://archive.org/details/principesdegogr00blacgoog website consulted in Février 2015.

In the annex 2, 10 lines of text are deconstructed at the 5th, 10th and 15th iteration. We took away the 27 most frequent characters in the text. The results of the deconstruction can be seen in table 1. Each line (except for the first one) corresponds to a step in the text's deconstruction (namely, the suppression of the characters in column 1). The first line corresponds to the text's characteristics before the deconstruction. It verifies Zipf's law with a coefficient close to 1, as in $\beta \approx 1.02$.

From left to right, we can read:
Column 1, characters deleted in the whole text
Column 2, the percentage of deleted characters in the text,
Column 3, S_i, the number of segmented sources,
Column 4, E_i, the amount of effort (see equation (6)),
Column 5, H_i, entropy (see equation (5)),
Column 6, the coefficient β_i (see equation (3)),
Column 7, k_i, the normalisation coefficient (see equation (3),
Colum 8, R^2, the squared linear coefficient

The second line (replacing « e » with @) is identical to the first line, since the structure of the text has not actually changed. At the 27th step, 79.5% of signs are deleted with approximately 1.5% of characters left. The only visible elements are the separators and the jokers @ (see the 15th deconstruction in the annex). The text is composed of 57 distinct characters. Capital letters have not yet been deleted. We must also remember that the percentage of separators remains constant throughout the entire deconstruction. The amount of effort and entropy decrease steadily at each step. We must check that the number of segmented sources S_i (see section 3.3) is a decreasing sequence and that β_i (the adjustment coefficient) is an increasing sequence[10], according to (10b) and (11). The normalisation coefficient increases steadily up to the 15th iteration. Variations then become irregular. We notice that the necessary and sufficient condition to solve equation (4) is verified. As an example, for the first deconstruction, we have:

$$M_1 = \frac{M}{k_1} = \frac{114{,}730}{15{,}754} = 7.28 \leq = \mathrm{Ln}(S_1) = \mathrm{Ln}(12{,}747) = 9.45$$

This calculation is verified at each step, with, at the 27th step:

$$M_{27} = \frac{M}{k_{27}} = \frac{114{,}730}{22{,}715} = 5.05 \leq = \mathrm{Ln}(S_{27}) = \mathrm{Ln}(936) = 6.84$$

We observe an acceptable adjustment for each source-frequency distribution up to the 27th iteration (936 sources). The coefficient of determination R^2 is equal to 0.97, which is the value of R^2 in the first adjustment before degrading the text. While the adjustment is not very good for high frequencies, it remains classic in form when we represent the distribution (frequency rank) of a text in general with a *Ln-Ln* scale. (see figure 1, figure 2)

[10] We can show that this sequence is decreasing thanks to the proposition proven in Annex 1 . Indeed, the S_i sequence decreases through construction and the $M_i - S_i$ sequence, which equals $\frac{680{,}54}{k_i} - S_i$ is also decreasing.

Car	%	S_i	E_i	H_i	β_i	k_i	R^2
		12747	1.02	6.77	1.05	15754	0.974
e	11.8	12747	1.02	6.77	1.05	15754	0.974
s	19.2	12481	1.02	6.76	1.05	16900	0.975
n	25.0	12429	1.01	6.72	1.05	16410	0.974
a	30.6	12155	1.00	6.66	1.06	17034	0.974
i	36.2	11821	0.99	6.62	1.07	18095	0.974
t	41.6	11400	0.98	6.54	1.08	19195	0.975
r	46.6	10513	0.96	6.43	1.10	22304	0.976
l	51.3	9848	0.91	6.20	1.11	23301	0.976
u	55.8	8991	0.85	5.93	1.13	24146	0.976
o	59.9	7807	0.80	5.68	1.16	28832	0.977
d	63.2	7149	0.74	5.34	1.18	30020	0.977
c	65.8	6194	0.69	5.09	1.22	34738	0.977
p	68.1	5420	0.64	4.84	1.24	36838	0.978
m	70.2	4495	0.06	4.58	1.29	41668	0.980
é	72.1	3307	0.54	4.28	1.39	60620	0.980
'	73.2	3297	0.52	4.12	1.38	56830	0.979
v	74.2	2916	0.48	3.90	1.38	49847	0.980
g	75.1	2457	0.45	3.70	1.41	50021	0.981
q	76.0	2257	0.42	3.55	1.43	48687	0.981
f	76.8	1974	0.40	3.37	1.44	44526	0.982
h	77.5	1611	0.38	3.22	1.50	47867	0.983
b	78.1	1349	0.35	3.04	1.51	39566	0.984
x	78.5	1242	0.34	2.93	1.50	32740	0.981
à	78.9	1235	0.34	2.91	1.49	30533	0.980
è	79.2	1142	0.33	2.83	1.48	24185	0.977
y	79.4	989	0.32	2.76	1.52	26168	0.975
j	79.5	936	0.32	2.72	1.51	22715	0.972

TABLE 1 – Results of the Deconstruction

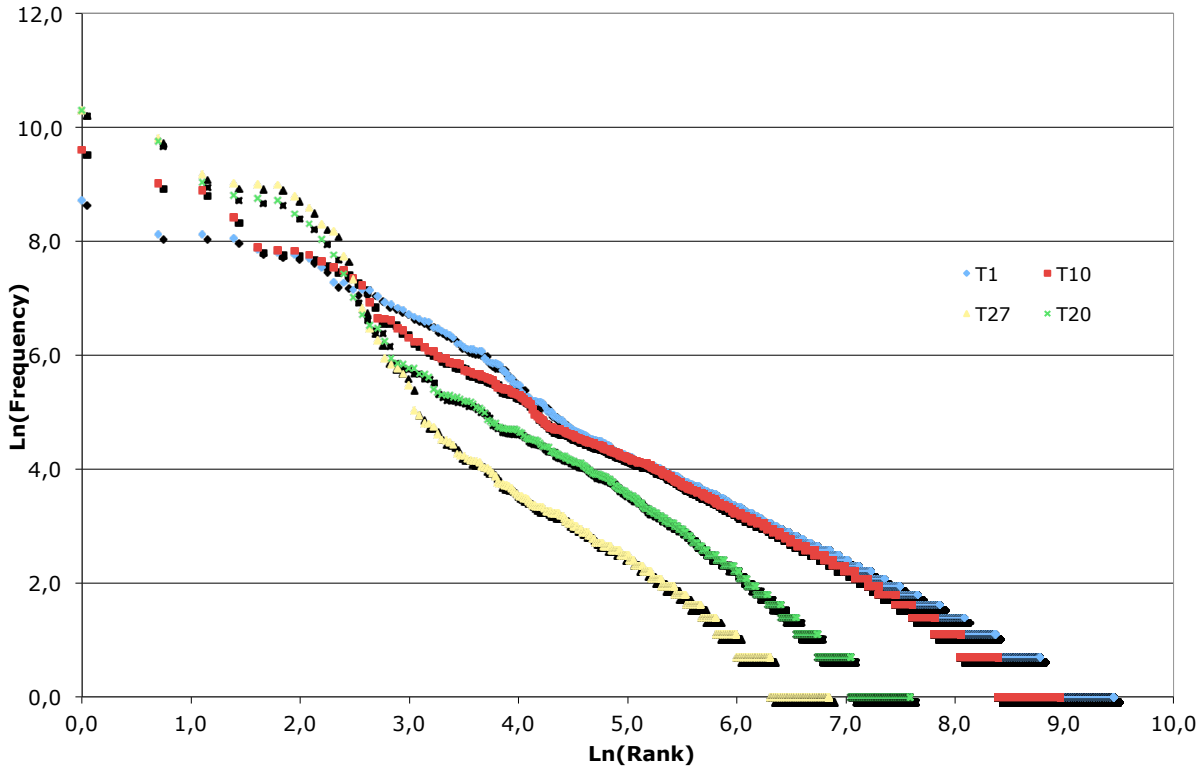


FIGURE 1 - The Source-frequency Distribution throughout the Deconstruction

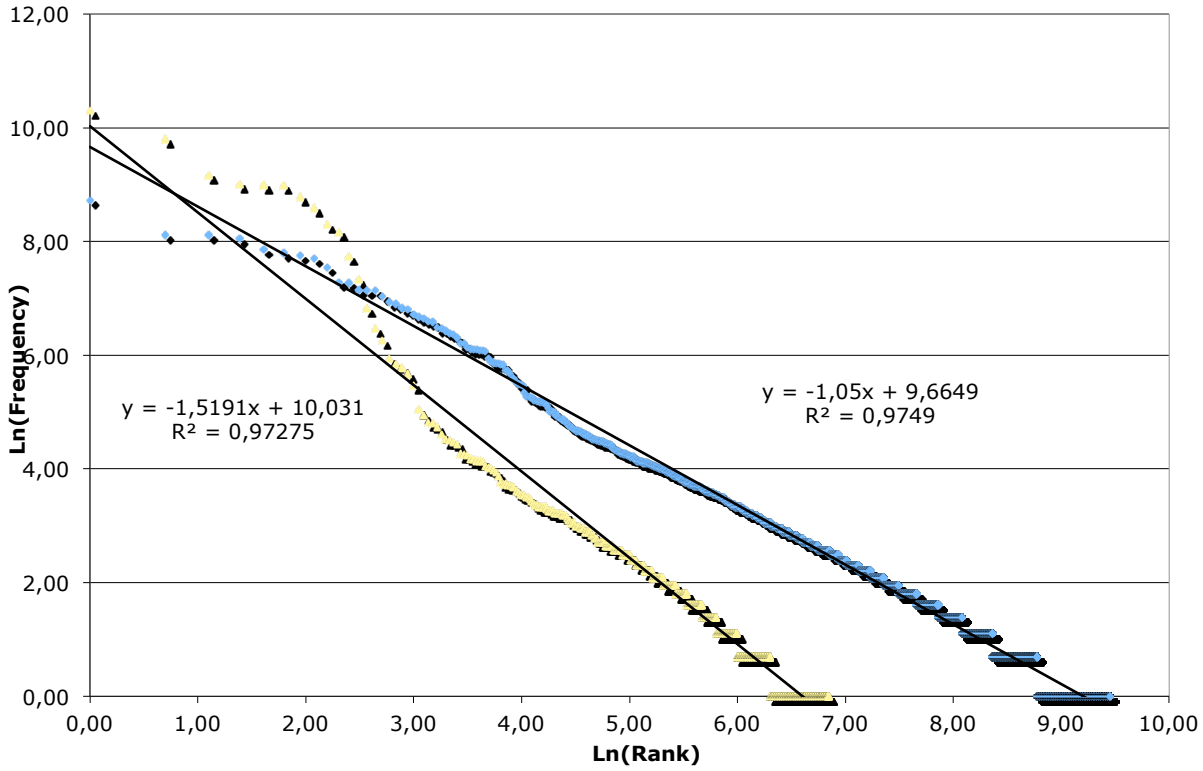


FIGURE 2 – The Adjustment of both Extreme Distributions

5.3 Other Deconstruction Experiments

We deconstructed our text again (this experiment is useful when studying the variation of entropy according to effort[11]). But this time, instead of deleting characters by decreasing frequency, we deleted them randomly. The initial point and the final point of deconstruction were the same. But the order of characters changed. Two experiments were conducted.
The results can be seen in Table 2
Column 1, 7: deleted characters
Column 2, 8 : number of segmented sources
Column 3,9 : effort amount
Column 4,10 : entropy
Column 5,11 : adjustment coefficient β.

Car	S_i	E_i	H_i	β_i		Car	S_i	E_i	H_i	β_i
x	12747	1.02	6.77	1.05		'	12747	1.02	6.77	1.05
m	12745	1.02	6.77	1.05		a	12743	1.01	6.73	1.04
é	12741	1.02	6.77	1.05		b	12743	1.01	6.73	1.04
à	12739	1.02	6.77	1.05		c	12723	1.01	6.73	1.04
è	12735	1.03	6.77	1.05		d	12685	1.00	6.68	1.04
r	12704	1.03	6.77	1.05		e	12404	0.97	6.54	1.04
p	12663	1.03	6.77	1.05		f	12362	0.98	6.54	1.04
d	12594	1.03	6.76	1.05		g	12306	0.98	6.53	1.04
c	12488	1.03	6.71	1.05		h	12270	0.98	6.53	1.04
o	12378	1.03	6.69	1.05		i	12049	0.97	6.49	1.05
u	12160	1.02	6.67	1.06		j	12015	0.98	6.48	1.05
l	11844	0.99	6.52	1.06		l	11742	0.93	6.28	1.05
e	10734	0.95	6.35	1.09		m	11452	0.93	6.25	1.06
j	10674	0.95	6.33	1.09		n	10613	0.90	6.09	1.08
y	10610	0.95	6.32	1.09		o	9795	0.87	5.95	1.11
s	9310	0.90	6.04	1.13		p	9186	0.86	5.85	1.12
n	8209	0.83	5.74	1.16		q	9134	0.86	5.84	1.12
t	6481	0.75	5.30	1.22		r	7084	0.80	5.57	1.2
i	4497	0.64	4.77	1.3		s	5161	0.70	5.04	1.29
a	2671	0.51	4.01	1.43		t	3248	0.57	4.38	1.41
'	2662	0.48	3.85	1.42		u	2204	0.44	3.65	1.48
b	2278	0.45	3.70	1.46		v	1805	0.41	3.42	1.51
h	1903	0.43	3.56	1.51		x	1652	0.39	3.31	1.53
f	1626	0.41	3.38	1.54		y	1467	0.38	3.26	1.57
q	1479	0.38	3.21	1.54		à	1459	0.38	3.23	1.56
g	1129	0.35	2.98	1.59		é	1021	0.33	2.8	1.53

[11] We could indeed assume that the observed results arose from the fact that the characters were deleted by decreasing frequencies.

v	936	0.32	2.72	1.51		è	936	0.32	2.72	1.51

Table 2: Lexical Results and Deconstruction Adjustments

We have the same properties as in previous cases: S_i decreases, we have an adjustment with the generalized Zipf law, β_i decreases. In the 2nd deconstruction, β_i's variation is more chaotic. Entropy and the amount of effort decrease steadily. In the next paragraph, we will see that the variation of these two quantities are linked.

6. Analysis and Discussion

6.1 Adjustment with the Generalized Zipf Law

We just saw that the adjustment is sustainable in all three experiments. Even after having been highly degraded, an adjustment with the generalized Zipf's law is still acceptable. We can't, however, consider the generalized Zipf law to be fully satisfactory. Figure 2 shows that the high frequencies stray away from the regression line. Such a deviation is common when one aims to verify Zipf's law. But here, the deformation is accentuated as the deconstruction occurs.

In the three experiments, 27 characters were deleted. Approximately 2% of the characters in the text remained. The text, degraded in this way, is no longer a text. The only things remaining throughout the deconstruction are the text's structure and the length of the words (see the text at the tenth step in Annex 2). β varies from 1.01 to 1.51; the adjustment coefficient would have had exactly the same variation if the texts had been randomly generated (see the use of formula (10b), section 4.2.4)..

If we were to delete all the characters, the text T_{83} would only be composed of words containing the @ joker. The final distribution (frequency rank) would correspond to the distribution of the length of the words. We know that this is not a generalized Zipf law.

If we had only found the adjustments as results, such a basic deconstruction approach wouldn't greatly contribute to real language and would simply be a novelty. In the following section, we therefore aim to quantify and connect the different states of the text.

6.2 The Linearity between Entropy and Amount of Effort.

In this section, we analyse the relationship between the entropy H_i and the amount of effort E_i obtained when deconstructing a text, in light of the theoretical results previously mentioned (see equations (11) (12). We notice a progressive decrease in the amount of effort E_i and in the entropy H_i.

After having done a linear regression on the pair of points (H_i, E_i) $i = 1..27$ (see table 1) we obtain, for the first experiment, the following equation:

$H_i \approx 5.61E_i + 1.1 \quad i = 1,..27 \quad R^2 \approx 0.996$ (see figure 3) (13)

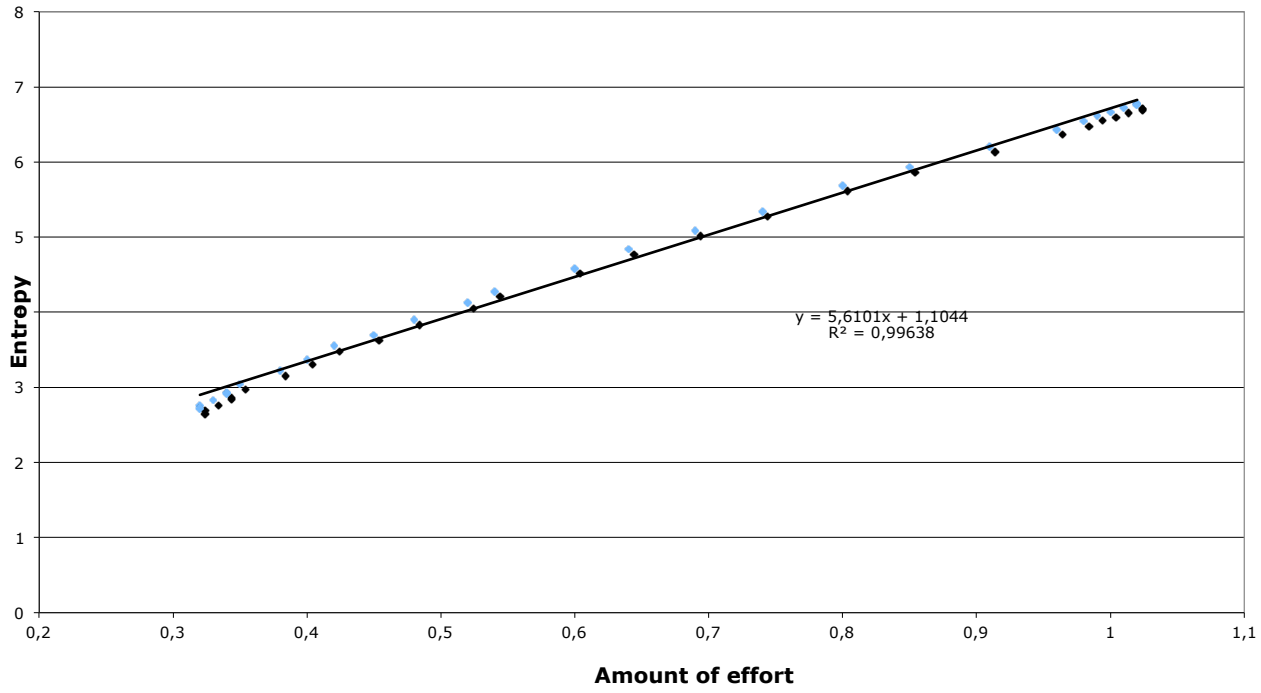


FIGURE 3 – Variations in Entropy according to the Amount of Effort

This result is surprising since there is an almost perfect linear relationship between entropy and the amount of effort. This result therefore differs from formula (12)

$$H_i = -\mathrm{Ln}(h_i) + \delta . E_i \qquad i = 1..I$$

Let us verify hypothesis (see p 9) $\beta_i . Ln(V_i)$ is constant and is equal to δ
If we calculate the average of the 27 values $\beta_i . Ln(V_i)$ (with V_i varying from 83 to 57) using table 2, we obtain 5.12 with a standard deviation of 0.70. This value is different from the 5.61 regression coefficient (see (13)) which was previously calculated.

In the two others deconstruction experiments, we obtain a near-perfect linear variation with the exception of the two extreme points, when the deconstruction is nearly complete:

$H_i \approx 5.45E_i + 1.16 \quad i = 1,..27 \quad R^2 \approx 0.998$ (see figure 4) (14)
$H_i \approx 5.56E_i + 1.1 \quad i = 1,..27 \quad R^2 \approx 0.998$ (see figure 4) (15)

By calculating the average of the values $\beta_i . \mathrm{Ln}(V_i)$ using the table 2, we obtain:
- 5.10 with a standard deviation of 0.74
- 5.19 with a standard deviation of 1.18
These values also differ from the regression coefficients, which are of 5.45 and 5.56.

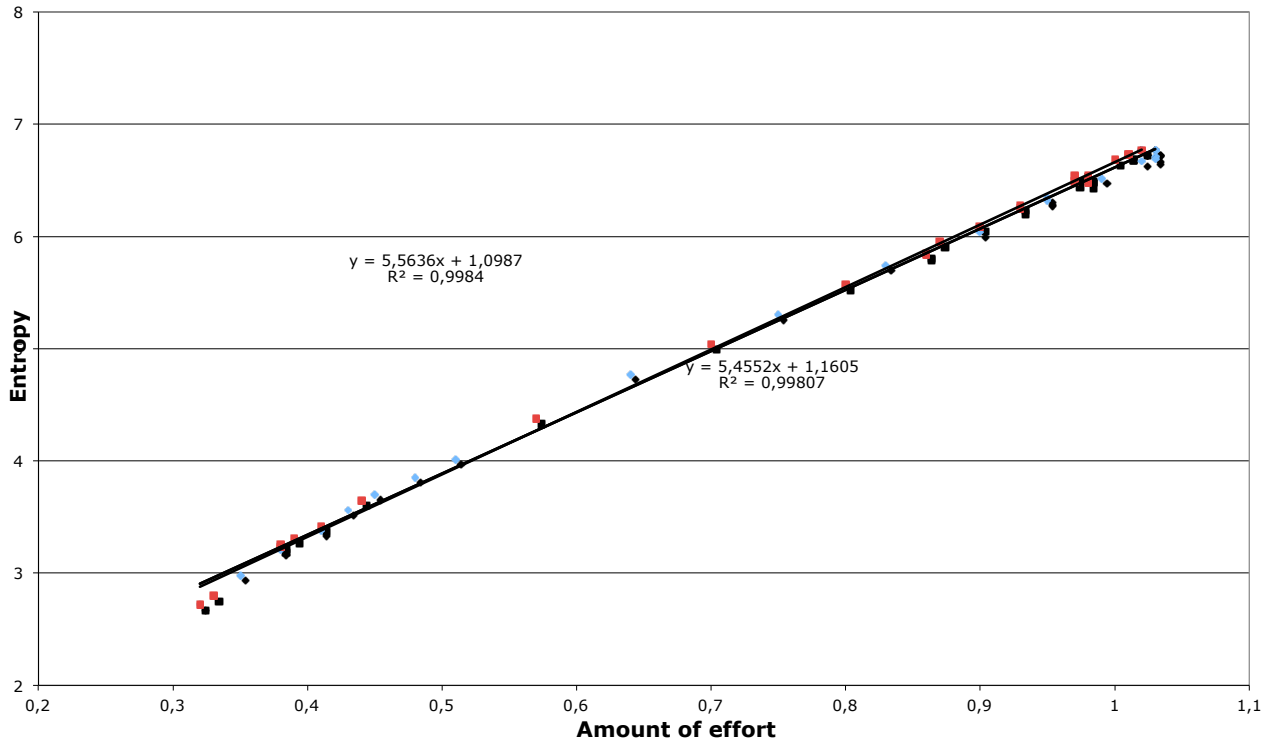


FIGURE 4 - Variation in Entropy according to the Quantity of Effort

Analysing and interpreting these results is delicate, mainly because the calculation of entropy only depends on segmentation and is the result of the experiment that is linked to deconstruction. Calculating the amount of effort depends on segmentation and the chosen effort function, as seen here:

$$C_i(r) = \frac{\mathrm{Ln}\,(r)}{\mathrm{Ln}\,(V_i)}$$

The observed linearity doesn't seem to depend on the adjustment parameters' results. We could have verified linearity (see (13),(14) (15)) directly by doing a regression, after having segmented the text and ranked the words by decreasing frequency. The product $\beta_i.\ln(V_i)$ is not rigorously constant. It depends on the adjustment coefficient's calculation. The obtained value is different from the regression coefficient's calculation. Such a difference could be due to a lack of precision when calculating the β coefficient, or, more largely, to the chosen model, which may not be accurate – or complex enough – throughout the deconstruction process.

6.3 Generalizing the Result

The results of the 3 experiments can be analyzed together. If we conduct a linear regression of the 77 pairs: $(H_i, E_i)\ i = 1..77$, where (H_i, E_i) are the entropy and the effort amount of the degraded text T_i, we obtain a quasi-perfect linear regression (after removing duplicates):

$$H_i \approx 5.49E_i + 1.16 \quad i = 1,..77 \quad R^2 \approx 0.997 \text{ (see figure 5)} \qquad (16)$$

We notice that, when the text is highly degraded (low entropy and low amount of efforts), the dots aren't completely aligned. However, we feel that this linearity isn't as questionable as the graph with the Zipf curve.

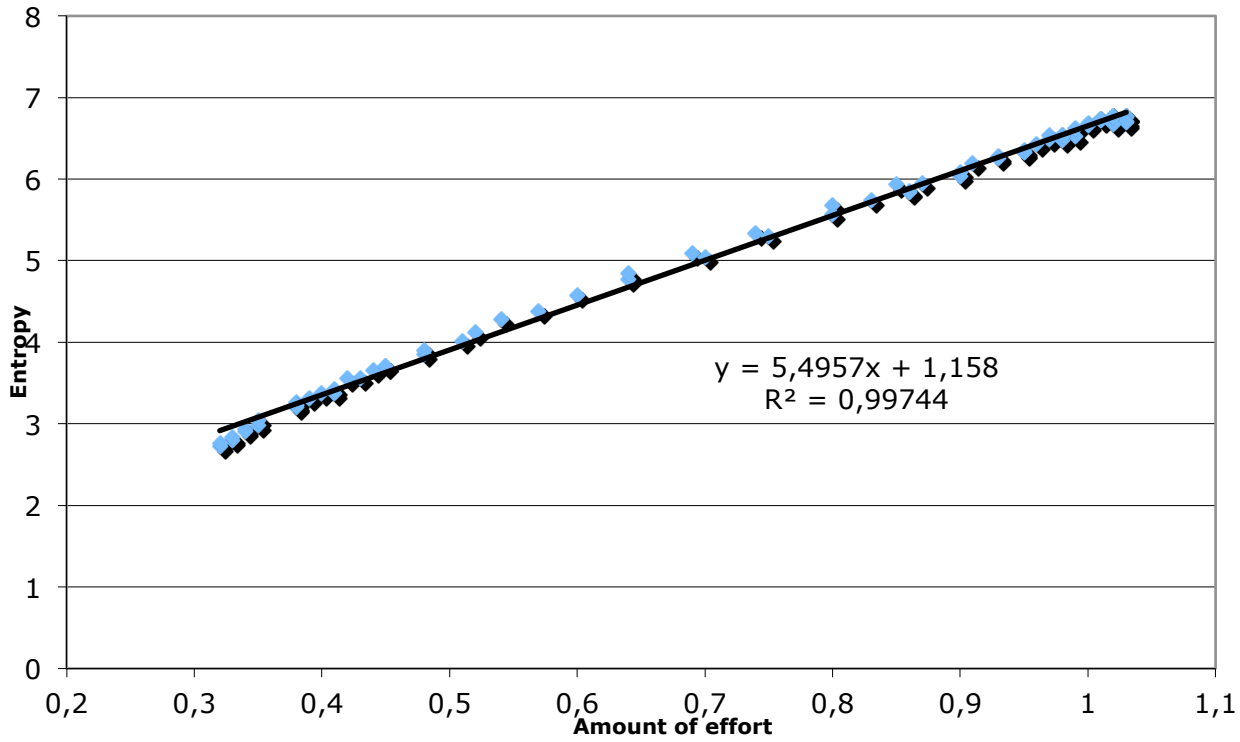


FIGURE 5 - Variation in Entropy according to the Quantity of Effort

Our hypotheses and the formula (12) – which we proved in section 4 with the help of Mandebrot's joint results combining entropy and amount of effort – are insufficient to explain the linearity observed in (13,14,15,16).
Let us first enounce the proven result in a more general way:

Let us call a text T and its set of distinct characters C (between 50 and 80 characters for a text written in French), thus $C_k \quad k = 1,..n$ is a sequence of n characters (between 20 and 30 following the deconstruction experiments), representing approximately 80% of the characters of the text. We then have the following result:

Let T_k be the text obtained after deleting from T the sequence of k characters $i = 1,..k \quad k \leq n$. We can then confirm that, on the basis of this study, if we call H_k the entropy and E_k the quantity of effort, there is a linear relationship of the type:

$$H_k = \gamma + \delta . E_k \qquad k = 1, \dots P \quad \delta > 0 \quad \gamma > 0 \tag{17}$$

Where P is the number[12] of deconstructed texts. In the example used, P is equal to 77.
In order to confirm the universality of such an inter-textual linearity, we present the degradation of another text in Annex 4. The deconstructed text is "The Value of Science" by Henri Poincaré, a philosophical essay of approximately 180 pages, published in 1901. A bibliometric analysis of this text has already been conducted in (Lafouge and Pouchot 2012), chapter 4, section 4.2.1.

We offer a possible explanation for this linearity with the following model:
More generally, if we assume that, at each step i, the source-frequency distribution can be written as:

[12] The number of texts that can be deconstructed is very high $\approx 10^7$!

$$p_i(r) = B.\exp\left(-A.F_i(r)\right) \quad A > 0 \quad i = 1..\ I \tag{18}$$

Where F_i is a sequence of functions (called effort function), *I* is the number of deconstructions. When we name, as we did before, H_i as entropy and E_i as the amount of effort, we can show (see Annex 3) that the entropy and amount of effort are linked by the relationship:

$$H_i = -\operatorname{Ln}(B) + A.E_i \tag{19}$$

In this case, the *A* and *B* constants are independent of *i*. Throughout the text's deconstruction, each step *i* is characterised by an effort function F_i.This formalism uses the effort function concept (Lafouge, Smolczewska, 2006) (Lafouge, Agouzal 2015).

If we choose $F_i(r) = \frac{Ln(r)}{Ln(V_i)}$, we find the previous result, which is $\beta_i = \frac{A}{\operatorname{Ln}(V_i)}$ meaning that A is constant. This model remains incomplete and isn't fully satisfactory, but provides us with new research perspectives.

6. Conclusion

This article's goal isn't to provide a new explanatory model for Zipf's law. However, the results that we obtained, and which could not be predicted, shed new light on the term "paradigm" used to qualify Zipf's law in the introduction. By degrading a text, we wanted to jointly explore both interpretations of *Mandelbrot*. What interested us was to confront both hypotheses. This naturally led us to focus on the entropy and the amount of effort of a degraded text. The reader may feel baffled by the frustrating method that consists in degrading a text mechanically with no regard for lexicographic knowledge (differentiating vowels from consonants…) or grammar (respecting the type of words: nouns, adjectives, conjunctions…). But this deconstruction can be applied to any text and, in this way, is a universal method. The only information used during the degradation process had to do with the status of the signs in the text as either being characters or separators. The number of steps for the algorithm corresponded roughly to the number of characters (between 20 and 30) that were essential to write the text. At the end of the deconstruction, only a few rare characters (in the statistical sense) were still visible. Wentiam Li's results (Li, W. (1992) are challenged by the fact that the generalized Zipf law applies to any deconstructed text. Li concludes his study by saying that "In conclusion, Zipf's law is not a deep law in natural language as one might first have thought". We could have thought that the adjustment wouldn't be valid after such a significant reduction of the number of sources (divided by 10 in our example).

The deconstruction has highlighted an unexpected result. If, like Mandelbrot (see 4.2.2), we define *C* as being the ratio between the amount of effort and entropy $C = \frac{E}{H}$, this article shows that, for the average cost of information – known as C_{ij} – the difference between two states of the same text, known as i and j, such that $\frac{E_i - E_j}{H_i - H_j}$, is constant. As in the original formulation of Zipf's law – namely that the rank by frequency is constant – the inter-textual linearity between entropy and effort seems to be a paradigm. This result puts the work of Mandelbrot in perspective concerning language generation and, more generally, concerning the models on commutative optimization principles. But we would need to confirm this by studying the variation range of the *C* coefficient on a larger amount of texts. We can however suppose that, since the adjustment coefficient β varies little and stays close to 1, the variation range of δ (see equation (17)) would not be very large. Can we say that it is another expression of the Zipf 's law, or a consequence of it?

Although the explanation given for equation (18) – an another formulation of the generalized Zipf law – is an approximation and isn't entirely satisfactory, it does, however, open new research perspectives. What happens at the end of the deconstruction, when the rare characters left also disappear?

We will end this brief conclusion with a warning by *Stumpf* (Stumpf and Porter, 2012) who analyzed the relevance and the validity of various studies on power laws in general and in various fields: "Power laws do have an interesting and possibly even important role to play, but one needs to be very cautious when interpreting them".

ANNEX 1

Theorem 4.1 Let a pair $(S_i, M_i) \in [1,..\infty[\times\mathbb{R}^+$, then there exists a unique real $x, x \geq 1$ such that:

$$\int_1^S \frac{dr}{r^x} = M$$

if and only if : $M \in [o,..\ln(S)]$

Proof

Let the function

$$F(x) = \int_1^x \frac{dr}{r^x}$$

where. $x \in [1..\infty[$. This function is strictly decreasing and continuous. Then we have:

$F(1) = ln\, s$.

For $x > 1$, we have:

$$F(x) = \frac{1}{x-1}.(1 - \frac{1}{S^{x-1}})$$

If x tends towards ∞ then $F(x)$ tends towards 0. The theorem of intermediate values then allows us to say that there is a unique value x such that $F(x) = M$.

Proposition We have the following results*:*

If S_i a decreasing sequence and M_i - S_i is a decreasing sequence, then β_i is an increasing sequence.

Proof

We show the identity:

$$Y = \int_1^{S_{i+1}} \left(\frac{1}{r^{\beta_{i+1}}} - \frac{1}{r^{\beta_i}}\right)dr = M_{i+1} - M_i + \int_{S_{i+1}}^{S_i} \frac{dr}{r^{\beta_i}}$$

we have also :

$$\int_{S_{i+1}}^{S_i} \frac{dr}{r^{\beta_i}} \leq \int_{S_{i+1}}^{S_i} dr = S_i - S_{i+1} < 0$$

S_i is decreasing and also $M_i - S_i$

hence, we can write:

$$Y \leq M_{i+1} - M_i + S_i - S_{i+1}$$

$$Y \leq (M_{i+1} - S_{i+1}) - (M_i - S_i) \leq 0$$

Then we have $Y < 0$ hence, we can write:

$$Y < 0 \Rightarrow \forall r \geq 1 \; \frac{1}{r^{\beta_{i+1}}} - \frac{1}{r^{\beta_i}} \leq 0$$

Thus

$$\forall r \geq 1 \; r^{\beta_i} \leq r^{\beta_{i+1}} \iff \beta_i \leq \beta_{i+1}$$

$$\boldsymbol{\beta_i \leq \beta_{i+1}}$$

ANNEX 2

ITERATION 1 : 12,747 sources

la géographi@ humain@ @st un@ d@s branch@s qui
ont réc@mm@nt poussé sur l@ vi@ux tronc d@ la
géographi@. s' il n@ s' agissait qu@ d' un@ épithèt@,
ri@n n@ s@rait moins nouv@au. l' élém@nt humain
fait @ss@nti@ll@m@nt parti@ d@ tout@ géographi@ ;
l' homm@ s' intér@ss@ surtout à son s@mblabl@, @t,
dès qu' a comm@ncé l' èr@ d@s pérégrinations @t
d@s voyag@s, c' @st l@ sp@ctacl@ d@s div@rsités
social@s associé à la div@rsité d@s li@ux qui a
piqué son att@ntion.

ITERATION 5 : 11,821 sources

l@ géogr@ph@@ hum@@@@ @@t u@@ d@@ br@@ch@@ qu@
o@t réc@mm@@t pou@@é @ur l@ v@@ux tro@c d@ l@
géogr@ph@@. @' @l @@ @' @g@@@@@t qu@ d' u@@ ép@thèt@,
r@@@ @@ @@r@@t mo@@@ @ouv@@u. l' élém@@t hum@@@
f@@t @@@@@t@@ll@m@@t p@rt@@ d@ tout@ géogr@ph@@ ;
l' homm@ @' @@tér@@@@ @urtout à @o@ @@mbl@bl@, @t,
dè@ qu' @ comm@@cé l' èr@ d@@ pérégr@@@t@o@@ @t
d@@ voy@g@@, c' @@t l@ @p@ct@cl@ d@@ d@v@r@@té@
@oc@@l@@ @@@oc@é à l@ d@v@r@@té d@@ l@@ux qu@ @
p@qué @o@ @tt@@t@o@

ITERATION 10 :7,807 sources

@@ gé@g@@ph@@ h@m@@@@ @@@ @@@ d@@ b@@@ch@@ q@@
@@@ @éc@mm@@@ p@@@@é @@@ @@ v@@@x @@@@c d@ @@
gé@g@@ph@@. @' @@ @@ @' @g@@@@@@ q@@ d' @@@ ép@@hè@@,
@@@@ @@ @@@@@@ m@@@@ @@@v@@@. @' é@ém@@@ h@m@@@
f@@@ @@@@@@@@@@@@m@@@ p@@@@@ d@ @@@@@ gé@g@@ph@@ ;
@' h@mm@ @' @@@é@@@@@ @@@@@@@@ à @@@ @@mb@@b@@, @@,
dè@ q@' @ c@mm@@cé @' è@@ d@@ pé@ég@@@@@@@@@@ @@
d@@ v@y@g@@, c' @@@ @@ @p@c@@c@@ d@@ d@v@@@@@é@
@@c@@@@@ @@@@c@é à @@ d@v@@@@@é d@@ @@@@x q@@ @
p@q@é @@@ @@@@@@@@@@.

ITERATION 15 : 3,307 sources

@@ g@@g@@@h@@ h@@@@@@ @@@ @@@ @@@ b@@@@h@@ q@@
@@@ @@@@@@@@@@ @@@@@@ @@@ @@ v@@@x @@@@@ @@ @@
g@@g@@@h@@. @' @@ @@ @' @g@@@@@@ q@@ @' @@@ @@@@hè@@,
@@@@ @@ @@@@@@ @@@@@ @@@v@@@. @' @@@@@@@ h@@@@@
f@@@ @@@@@@@@@@@@@@@@ @@@@@@ @@ @@@@@ g@@g@@@h@@ ;
@' h@@@@ @' @@@@@@@@@@ @@@@@@@ à @@@ @@@b@@b@@, @@,
@è@ q@' @ @@@@@@@@@ @' è@@ @@@ @@@@g@@@@@@@@@@ @@
@@@ v@y@g@@, @' @@@ @@ @@@@@@@@@@ @@@ @@v@@@@@@@
@@@@@@@@ @@@@@@@@ à @@ @@v@@@@@@ @@@ @@@@x q@@ @
@@q@@ @@@ @@@@@@@@@@.

ANNEX 3

Demonstration of the relationship between the amount of effort and entropy (9) in paragraph 4.2.3

If we calculate entropy $H_i = -\sum_{r=1}^{S_i} \mathrm{Ln}\,(p_i\,(r)).p_i(r)$ assuming that the distributions are generalized Zipf law, using the equation $f_i(r) \approx \frac{k_i}{r^{\beta_i}}\ \ k_i > 0\ \ \beta_i \geq 1\ \ r = 1,..S_i$ with $h_i = \frac{k_i}{M_i}$ and $\int_1^{S_i} \frac{dr}{r^{\beta_i}} = M_i\ \ ;\quad M_i = \frac{M}{k_i}$ we have:

$$H_i = -\,\mathrm{Ln}(h_i) + \beta_i \sum_{r=1}^{S_i} \frac{h_i}{r^{\beta_i}} \mathrm{Ln}(r) \qquad \beta_i\ \geq 1, 0 < h_i$$

according to $C_i(r) = \frac{\mathrm{Ln}\,(r)}{\mathrm{Ln}\,(V_i)}$ we have:

$$H_i = -\,\mathrm{Ln}(h_i) + \beta_i.\mathrm{Ln}(V_i)\,.\sum_{r=1}^{S_i} \frac{h_i}{r^{\beta^i}}\ .C_i(r)$$

according to $E_i = -\sum_{r=1}^{S_i} C_i\,(r).p_i(r)$ we have:

$$H_i = -\,\mathrm{Ln}(h_i) + \beta_i.\mathrm{Ln}(V_i)\,.E_i$$

we obtain a relationship between entropy and amount of effort:

$$H_i = \gamma_i + \delta_i.E_i \quad \gamma_i > 0, \quad \delta_i = \beta_i.\mathrm{Ln}(V_i) > 0 \qquad (9)$$

Demonstration of the relationship (16) between the amount of effort and entropy in paragraph 6.3

$p_i(r) = B.\exp\big(-A.F_i(r)\big)\ \ A > O\ \ i = 1..\ I$ where F_i is a strictly increasing unbounded effort function we suppose :

$$\sum_{i=1}^{I} p_i\,(r) = 1$$

then $H_i = -\ln(B) + A.E_i$

Proof

$$H_i = \sum_{i=1}^{I} Ln(\,p_i(r)).p_i(r)$$

$$H_i = -\sum_{i=1}^{I} Ln(B.\exp\big(-A.F_i(r)\big)\,.B.\exp\big(-A.F_i(r)\big)$$

$$H_i = -\sum_{i=1}^{I} Ln(B).B.\exp\big(-A.F_i(r)\big) - \sum_{i=1}^{I} -A.F_i(r).B.\exp\,(-A.F_i(r))$$

$$H_i = -\sum_{i=1}^{I} Ln(B).p_i\,(r) + \sum_{i=1}^{I} A.F_i(r).p_i(r)$$

$$H_i = -\ln(B)\sum_{i=1}^{I}.p_i\,(r) + A.\sum_{i=1}^{I}F_i\,(r).p_i(r)$$

$$E_i = \sum_{i=1}^{I}F_i\,(r).p_i(r)$$

$$H_i = -\ln(B) + A.E_i$$

If $F_i(r) = \frac{Ln(r)}{Ln(V_i)}$ is the effort function defined by *Mandelbrot* :

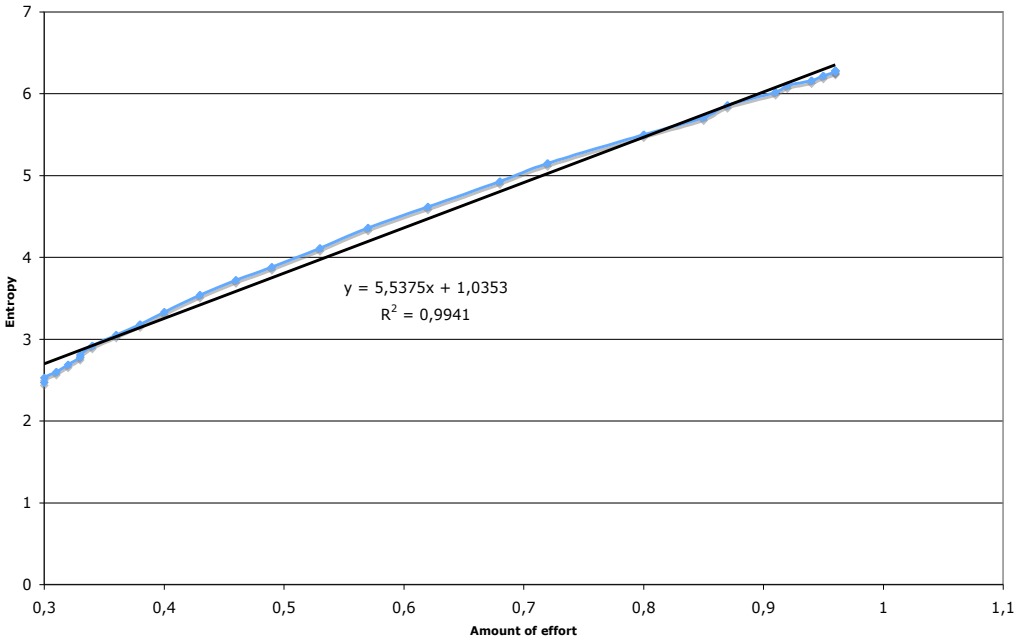
$$p_i(r) = B.\exp\left(-A.\frac{Ln(r)}{V_i}\right) = B.\frac{1}{r^{\beta_i}} \quad \beta_i = \frac{A}{\ln\,(V_i)}$$

ANNEX 4

The text's Characteristic

Total number of signs : 344,292
Percentage of separators : 20,2%
Percentage of character : 79,8 %
Total number of degmented words : 61,086
Number of identified sources : 5,853
Number of distinct character : 75
Adjustment coefficient β:1,12

This text is available on line using the URL http://fr.wikisource. org/wiki/ La_Valeur_de_la_Science, website consulted in February 2015.



Car	E	H
	0.96	6.28
e	0.96	6.28
s	0.96	6.27
n	0.95	6.22
t	0.94	6.16
i	0.92	6.1
a	0.91	6.02
u	0.87	5.86
r	0.85	5.71
o	0.8	5.5
l	0.72	5.15
c	0.68	4.93
d	0.62	4.62
p	0.57	4.36
m	0.53	4.11
é	0.49	3.88
'	0.46	3.72
q	0.43	3.54
v	0.4	3.33
f	0.38	3.18
b	0.36	3.05
g	0.34	2.92
h	0.33	2.81
à	0.33	2.78
x	0.32	2.69
j	0.31	2.6
è	0.3	2.53
y	0.3	2.47

Variations in Entropy according to the Amount of Effort

Result of deconstruction at the 4^{th} and 12^{th} sources)iteration

la recherche de la vérité doit être le but de
notre activité ; c' est la seule fin qui soit digne
d' elle. sans doute nous devons d' abord nous efforcer
de soulager les souffrances humaines, mais pourquoi ?

5,308 sources $\beta_4 = 1{,}17$

la r@ch@rch@ d@ la véri@é doi@ ê@r@ l@ bu@ d@
@o@r@ ac@ivi@é ; c' @@@ la @@ul@ fi@ qui @oi@ dig@@
d' @ll@. @a@@ dou@@ @ou@ d@vo@@ d' abord @ou@ @fforc@r
d@ @oulag@r l@@ @ouffra@c@@ humai@@@, mai@ pourquoi ?

1,811 sources $\beta_{12} = 1{,}31$

@@ @@@h@@@h@ d@ @@ vé@@@é d@@@ ê@@@ @@ b@@ d@
@@@@@@ @@@@v@@é ; @' @@@ @@ @@@@@@ f@@ q@@ @@@@ d@g@@
d' @@@@. @@@@ d@@@@ @@@@ d@v@@@ d' @b@@d @@@@ @ff@@@@@
d@ @@@@@@g@@ @@@ @@@ff@@@@@@@ h@m@@@@@m@@p@@@q@@@ ?

References

Benguigui, L. and Blumenfeld-Lieberthal, E. (2011). The end of a paradigm : is Zipf's law universal. *J Geogr Syst,* 281 :69-77

Clauset, A., Slalizi, C., and Newman, M. (2009). Power-law distribution empirical data. *SIAM Reviews*, 51 :661–771.

Egghe, L. (1990). On the duality of informetric systems with application to the empirical law. *Journal of Information Science,* 16 :17-27

Egghe, L. (2005). Power Laws in the Information Production Process: Lotkaian Informetrics. Elsevier.

Egghe, L. (2013). The functional relation between the impact factor and the uncitedness factor revisited. *Journal of Informetrics, 7* :183-189

Estoup, J. (1916). *Gammes sténographiques*. 4° édition, Institut Sténographique, Paris

Ferrer i Cancho, R, and Elveag, B. (2010). Random texts do not exhibit the real Zipf's law-like rank distribution. *Plos ONE,* 3 :1-9

Lafouge, T. and Pouchot, S. (2012). *Statistiques de l'intellect : lois puissances inverses en sciences humaines et sociales.* Publibook. Sous la direction de E. Guichard.

Lafouge, T. and Smolczewska, A. (2006). An interpretation of the effort function through the mathematical formalism of exponential informetric process. *Information Processing and Management,* 42 :1442-1450

Lafouge, T. and Agouzal, A. (2015). The Source-effort Coverage of an Exponential Informetric Process. Journal of Informetrics, 9 (1) :156-168.

Li, W. (1992) Random texts exhibits Zipf's law like word frequency distribution. *Information Theory, IEEE Transactions on*, 38(6), 1842-1845.

Lotka, A. (1926). The frequency distribution of scientific productivity. *Journal of the Washington Academy of Science,* 16 :317-323.

Mandelbrot, B. (1953). *An informational Theory of the statistical Structure of languages,* chapter Communication Theory, pages 486-502. W. Jackson. Woburn, MA : Butterworth.

Mandelbrot, B. (1977). *The Fractal Geometry of nature.* Freeman, New York USA

Mitzenmacher, M. (2003). A brief history of generative models for power law and lognormal distribution. *Internet Mathematics*, 1(2) :226-251

Petruszewycz, M. (1972). Loi de Pareto ou loi log-normale : un choix difficile. *Mathématiques et sciences humaines*, 39 :37–52.

Petruszewycz, M. (1973). L'histoire de la loi d' Estoup-Zipf. *Mathématiques des sciences humaines,* 11(44) :41-56

Piantadosi, Steven T**. (2014) . Zipf's word frequency law in natural language : a critical rewiew and future directions.** *Psychonomic Bulletin & Review* **21 (5) :1112-1130**

Price, D. D. S. (1976). A general theory of bibliometric and other cumulative and other advantage processes. *Journal of the American Society for Information Science,* 27(5-6) :292-306

Reginald, S. and Bouchet-Franklin Institute (2007). Investigation of the Zipf-plot of the extinct Meroitic language. *Glottometrics* 15 :53-61.

Simon, H. A. (1955). On a class of skew distributions functions. *Biometrika,*42(3/4) :425-440

Stumpf Michael P. H and Porter Masson A. (2012). Critical Truths About Power Laws. Sciences Vol 335, 665-666

Zipf, G. (1949). *Human behavior and the principle of least effort.* Cambridge, MA, USA Addison-Wesley. Reprinted : Hafner, New York, USA,1965

,